# LAB 11

*STAT 28*

*April 13, 2017*

Welcome to the lab 11! Today, we will use linear regression to predict the red wine quality using physicochemical tests scores such as citric acid, pH, etc.

The dataset is related to red variants of the Portuguese "Vinho Verde" wine. There are 1599 samples available in the dataset. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

The explanatory variables are all continuous variables based on physicochemical tests:

- **fixed acidity**
- **volatile acidity**
- **citric acid**
- **residual sugar**
- **chlorides**
- **free sulfur dioxide**
- **total sulfur dioxide**
- **density**
- **pH**
- **sulphates**
- **alcohol**

The response variable is the **quality** score between 0 and 10 (based on sensory data).

Read data. We randomly split the data into two parts-the `wine` dataset with 1199 samples and the `wine.test` dataset with 400 samples. Splitting the dataset is a common technique when we want to evaluate the model performance. There are training set, validation set, and test set. The validation set is used for model selection. That is, to estimate the performance of the different model in order to choose the best one. The test set is used for estimating the performance of our final model.

```
set.seed("20170413")
wine.dataset <- read.csv("winequality-red.csv", sep = ";")
test.samples <- sample(1:nrow(wine.dataset), 400)
wine <- wine.dataset[-test.samples, ]
wine.test <- wine.dataset[test.samples, ]
```

To check the correlation between explanatory variables:

```
cor(wine[, -1])
```

```
##                      volatile.acidity citric.acid residual.sugar
## volatile.acidity         1.0000000000 -0.55127829   -0.014100363
## citric.acid             -0.5512782908  1.00000000    0.149370933
## residual.sugar          -0.0141003633  0.14937093    1.000000000
## chlorides                0.0551519804  0.21471583    0.059074431
## free.sulfur.dioxide     -0.0006092232 -0.06888512    0.193370984
## total.sulfur.dioxide     0.0904268106  0.01469970    0.192606098
## density                  0.0079593988  0.38069677    0.353436849
## pH                       0.2495287298 -0.54800955   -0.078705732
## sulphates               -0.2533752753  0.31625615   -0.004154683
## alcohol                 -0.2003830982  0.11367551    0.048951341
## quality                 -0.3828352698  0.22616872    0.013948576
```

```
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## volatile.acidity      0.05515198       -0.0006092232           0.09042681
## citric.acid           0.21471583       -0.0688851240           0.01469970
## residual.sugar        0.05907443        0.1933709844           0.19260610
## chlorides             1.00000000        0.0177137828           0.04690574
## free.sulfur.dioxide   0.01771378        1.0000000000           0.66283723
## total.sulfur.dioxide  0.04690574        0.6628372272           1.00000000
## density               0.20430366       -0.0083136995           0.08273940
## pH                   -0.26818457        0.0659330393          -0.05523912
## sulphates             0.38414982        0.0359377655           0.04017958
## alcohol              -0.21427358       -0.0707444146          -0.21115760
## quality              -0.12673717       -0.0610430838          -0.19263493
##                          density          pH    sulphates     alcohol
## volatile.acidity      0.007959399  0.24952873 -0.253375275 -0.20038310
## citric.acid           0.380696768 -0.54800955  0.316256147  0.11367551
## residual.sugar        0.353436849 -0.07870573 -0.004154683  0.04895134
## chlorides             0.204303656 -0.26818457  0.384149822 -0.21427358
## free.sulfur.dioxide  -0.008313699  0.06593304  0.035937766 -0.07074441
## total.sulfur.dioxide  0.082739398 -0.05523912  0.040179577 -0.21115760
## density               1.000000000 -0.34763486  0.152160372 -0.48859800
## pH                   -0.347634862  1.00000000 -0.233160315  0.20155430
## sulphates             0.152160372 -0.23316032  1.000000000  0.09399430
## alcohol              -0.488598004  0.20155430  0.093994303  1.00000000
## quality              -0.176462068 -0.06812987  0.244380785  0.48507097
##                          quality
## volatile.acidity     -0.38283527
## citric.acid           0.22616872
## residual.sugar        0.01394858
## chlorides            -0.12673717
## free.sulfur.dioxide  -0.06104308
## total.sulfur.dioxide -0.19263493
## density              -0.17646207
## pH                   -0.06812987
## sulphates             0.24438079
## alcohol               0.48507097
## quality               1.00000000
```

Great! The correlations are not as high as the diamond dataset we saw in the last lab, which means we do not need to worry too much about heteroscedasticity. We now fit the linear regression using all of the explanatory variables:

```
wine.fit <- lm(quality ~. ,data = na.omit(wine))
summary(wine.fit)
```

```
##
## Call:
## lm(formula = quality ~ ., data = na.omit(wine))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64407 -0.34797 -0.05073  0.45560  2.01919
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.196e+01  2.454e+01   0.895 0.370963
```

```
## fixed.acidity         2.488e-02  3.002e-02   0.829 0.407530
## volatile.acidity     -1.040e+00  1.392e-01  -7.468 1.57e-13 ***
## citric.acid          -1.795e-01  1.703e-01  -1.055 0.291850
## residual.sugar        1.292e-02  1.710e-02   0.756 0.450065
## chlorides            -1.873e+00  4.798e-01  -3.903 0.000101 ***
## free.sulfur.dioxide   3.851e-03  2.510e-03   1.534 0.125237
## total.sulfur.dioxide -3.145e-03  8.452e-04  -3.721 0.000208 ***
## density              -1.788e+01  2.506e+01  -0.714 0.475512
## pH                   -4.278e-01  2.211e-01  -1.935 0.053259 .
## sulphates             8.579e-01  1.285e-01   6.676 3.75e-11 ***
## alcohol               2.820e-01  3.053e-02   9.236  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6483 on 1187 degrees of freedom
## Multiple R-squared:  0.3628, Adjusted R-squared:  0.3569
## F-statistic: 61.44 on 11 and 1187 DF,  p-value: < 2.2e-16
```

**Exercise 1**

As you can read from the p-values, several coefficients of the explanatory variables are not significant. Suppose now that we want to test the hypothesis that the betas corresponding to *fixed.acidity*, *citric.acid*, *residual.sugar* and *density* are all simultaneously zero.

(a) Fit the regression corresponds to the above submodel. Print the summary of your fit.

```
# insert your code here and save your fit as `submodel.fit`
# submodel.fit <-
# summary(submodel.fit)
```

(b) What is the degree of freedoms corresponding to your full model and your submodel?

```
# insert your code here.
# save the degree of freedom of your full model as `df.fullmodel`.
# save the degree of freedom of your submodel as `df.submodel`.
# df.fullmodel <-
# df.fullmodel
# df.submodel <-
# df.submodel
```

(c) Calculate the RSS (residual sum of squares) of the full model and submodel.

```
# insert your code here.
# save the RSS of your full model as `rss.fullmodel`.
# save the RSS of your submodel as `rss.submodel`.
# rss.fullmodel <-
# rss.fullmodel
# rss.submodel <-
# rss.submodel
```

(d) Calculate the F-statistics.

```
# insert your code here and save the F-statistics as `f.stat`
# f.stat <-
# f.stat
```

(e) Based on your F-statistics, calculate the p-value. Will you accept the full model or the submodel?

```
# insert your code here and save the p-value as `p.value`
# p.value <-
# p.value
```

(f) Use the `anova` function to verify your calculation above.

```
# insert your code here
```

**Exercise 2 Confidence Interval**

(a) Calculate the confidence interval for all the coefficients in the `submodel.fit`. Which of these factors will positively influence the wine quality?

```
# Insert your code here to calculate the confidence intervals for the regression coefficients.
```

(b) Calculate the confidence intervals for the samples in `wine.test` using `submodel.fit`. Which confidence interval will you use? Confidence intervals for the average response or the prediction interval?

```
# insert your code here and save your confidence intervals as `wine.confint`
# wine.confint <-
```

(c) What is the percentage that your interval in (b) covers the true **quality** score? What if you use the other confidence interval? Which one is consistent with your confidence level?

```
# insert your code here and save your percentage as `pct.covered`
# pct.covered <-
# pct.covered
# insert your code here and save your percentage calculated
# using the other confidence interval as `pct.covered.other`
# wine.confint.other <-
# pct.covered.other <-
# pct.covered.other
```

**Exercise 3 Backward Elimination based on p-values**

We start with our full model `wine.fit`.

(a) Remove the term with the highest p-value in the full model. Print the summary of your updated model.

```
# insert your code here and save your updated model as `wine.backward`
# wine.backward <-
# summary(wine.backward)
```

(b) Repeat the process in (a) until all the p-values are less than the significance level. Print the summary of your final model.

```
# insert your code here and save your model as `wine.backward`
```

(c) In R, there are functions which help us to do the automatically variable selection. The `step` uses AIC, criteria very similar to RSS but also takes the number of explanatory variables into account, to do the stepwise variable selection. For example, to do backward elimination starting with our full model:

```
step(wine.fit, direction = "backward")
```

```
## Start:  AIC=-1027.52
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                       Df Sum of Sq    RSS      AIC
## - density              1     0.214 499.03 -1029.01
```

```
## - residual.sugar        1     0.240 499.06 -1028.95
## - fixed.acidity         1     0.288 499.11 -1028.83
## - citric.acid           1     0.467 499.29 -1028.40
## <none>                              498.82 -1027.52
## - free.sulfur.dioxide   1     0.989 499.81 -1027.15
## - pH                    1     1.573 500.39 -1025.75
## - total.sulfur.dioxide  1     5.819 504.64 -1015.61
## - chlorides             1     6.400 505.22 -1014.24
## - sulphates             1    18.731 517.55  -985.32
## - volatile.acidity      1    23.439 522.26  -974.47
## - alcohol               1    35.845 534.66  -946.32
##
## Step:  AIC=-1029.01
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS      AIC
## - residual.sugar        1     0.070 499.10 -1030.84
## - fixed.acidity         1     0.078 499.11 -1030.82
## - citric.acid           1     0.474 499.51 -1029.87
## <none>                              499.03 -1029.01
## - free.sulfur.dioxide   1     1.088 500.12 -1028.40
## - pH                    1     3.537 502.57 -1022.54
## - total.sulfur.dioxide  1     6.085 505.12 -1016.48
## - chlorides             1     6.698 505.73 -1015.02
## - sulphates             1    18.906 517.94  -986.42
## - volatile.acidity      1    24.477 523.51  -973.59
## - alcohol               1    95.375 594.41  -821.31
##
## Step:  AIC=-1030.84
## quality ~ fixed.acidity + volatile.acidity + citric.acid + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol
##
##                        Df Sum of Sq    RSS      AIC
## - fixed.acidity         1     0.088 499.19 -1032.63
## - citric.acid           1     0.448 499.55 -1031.76
## <none>                              499.10 -1030.84
## - free.sulfur.dioxide   1     1.161 500.26 -1030.05
## - pH                    1     3.518 502.62 -1024.42
## - total.sulfur.dioxide  1     6.016 505.12 -1018.47
## - chlorides             1     6.638 505.74 -1017.00
## - sulphates             1    18.849 517.95  -988.39
## - volatile.acidity      1    24.407 523.51  -975.59
## - alcohol               1    96.874 595.98  -820.15
##
## Step:  AIC=-1032.63
## quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS      AIC
## - citric.acid           1     0.369 499.56 -1033.74
## <none>                              499.19 -1032.63
```

```
## - free.sulfur.dioxide   1     1.230 500.42 -1031.68
## - pH                    1     5.641 504.83 -1021.15
## - total.sulfur.dioxide  1     6.844 506.04 -1018.30
## - chlorides             1     7.525 506.72 -1016.69
## - sulphates             1    19.126 518.32  -989.55
## - volatile.acidity      1    25.205 524.40  -975.57
## - alcohol               1    97.536 596.73  -820.64
##
## Step:  AIC=-1033.74
## quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                         Df Sum of Sq    RSS      AIC
## <none>                               499.56 -1033.74
## - free.sulfur.dioxide   1     1.448 501.01 -1032.27
## - pH                    1     5.601 505.16 -1022.37
## - total.sulfur.dioxide  1     7.480 507.04 -1017.92
## - chlorides             1     8.385 507.95 -1015.78
## - sulphates             1    18.898 518.46  -991.22
## - volatile.acidity      1    29.338 528.90  -967.32
## - alcohol               1    99.225 598.79  -818.51
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = na.omit(wine))
##
## Coefficients:
##          (Intercept)      volatile.acidity             chlorides
##             4.411327             -0.973184             -2.025757
##  free.sulfur.dioxide  total.sulfur.dioxide                    pH
##             0.004549             -0.003380             -0.493988
##            sulphates               alcohol
##             0.828721              0.294570
```

Now try to understand the output of `step` function. Write down the order of variable elimination and the final model.

Order of variable elimination:
Final model:

(d) Start from the model with only intercept term. Use the step function to do the forward selection. Write down the order of variable addition and the final model. HINT: (a) Use the `scope` argument in step function. (b) Use `formula` function to get the formula of your full model.

```
# Insert your code here
```

Order of variable addition:
Final model: